PFID: PITTSBURGH FAST-FOOD IMAGE DATASET

Mei Chen[1], Kapil Dhingra[3], Wen Wu[2], Lei Yang[2], Rahul Sukthankar[1], Jie Yang[2]
[1]Intel Labs Pittsburgh, [2]Carnegie Mellon University, [3]Columbia University
http://pfid.intel-research.net

ABSTRACT

We introduce the first visual dataset of fast foods with a total of 4,545 still images, 606 stereo pairs, 303 360^0 videos for structure from motion, and 27 privacy-preserving videos of eating events of volunteers. This work was motivated by research on fast food recognition for dietary assessment. The data was collected by obtaining three instances of 101 foods from 11 popular fast food chains, and capturing images and videos in both restaurant conditions and a controlled lab setting. We benchmark the dataset using two standard approaches, color histogram and bag of SIFT features in conjunction with a discriminative classifier. Our dataset and the benchmarks are designed to stimulate research in this area and will be released freely to the research community.

Index Terms**—** Food image dataset, object recognition

1. INTRODUCTION

Image datasets are a prerequisite to visual object recognition research such as object modeling, detection, classification, and recognition. In fact, publicly available data collection and evaluation play a vital role in the development of automated object recognition technologies. The research community has developed both general- and specific-purpose datasets. The former contains a variety of objects and is primarily designed to support category-level object recognition research (e.g., the TU Darmstadt Dataset [1], Caltech 101 [3], Caltech 256 [13], the VOC2005 [4], and VOC2008 [5]). By providing standardized data on which researchers can train and test their algorithms, such datasets have made it possible to compare different approaches for object category recognition, and algorithm development has been spurred by large-scale competitions such as the PASCAL VOC 2008 object recognition challenge. Specific purpose image datasets, on the other hand, typically serve to accelerate research in a particular area. For instance, the face recognition community has benefited from a series of U.S. Government funded technology development efforts and evaluation cycles, beginning with the Facial Recognition Technology (FERET) program in 1993 [6]. The evaluations have documented two orders of magnitude improvement in performance from the start of the FERET program through the Face Recognition Vendor Test (FRVT) in 2006. In this effort, our intent is to stimulate research in automated food recognition by providing the research community with a comprehensive, public dataset of common fast food items acquired under both controlled and natural conditions.

Currently, there is no public dataset dedicated to automated visual food recognition. Automated food recognition is a key technology for measuring dietary and supplement intake in obesity study and treatment. Accurate and passive acquisition of dietary data from free-living individuals is essential for a better understanding of the etiology of obesity and the development of effective weight management programs. Currently, self-reporting is the main method for data acquisition. Despite its wide application through the use of questionnaires and structured interviews, numerous studies have revealed that data obtained by self-reporting seriously underestimate food intake, and thus do not accurately reflect the habitual behavior of individuals in real life. Our previous research [2,7,8] has proposed the use of computer vision to improve the accuracy of food intake reporting.

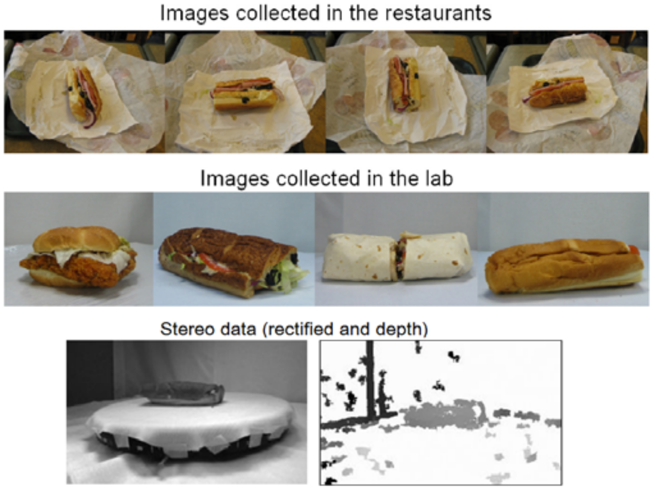
Images collected in the restaurants
Images collected in the lab
Stereo data (rectified and depth)

Figure 1: Examples from the Pittsburgh Fast-Food Image Dataset.

This paper presents the Pittsburgh Fast-food Image Dataset (PFID) (Figure 1), a collection of visual data to facilitate research in automated food recognition. PFID contains data of 101 fast food acquired in both restaurant environments and laboratory settings. It offers still images, video, and stereo images to support different algorithm developments and evaluations. We provide baseline evaluation results on food recognition. The dataset and evaluations are freely available to the public at http://pfid.intel-research.net.

2. DATA COLLECTION

We focus on fast food because it is standardized in terms of ingredients and preparation, and is seen as key to obesity research.

2.1. Select fast food chains

We identified eleven popular fast food chains, most of which are in USDA Food and Nutrient Database for Dietary Studies [14] and National Nutrient Databases. We selected 101 different foods from these chains including burgers, pizza, burgers, salads, etc.

2.2. Collection protocol and equipment

We collected three instances of each food item, which were purchased on different dates from the same restaurant or at different branches of the same fast food chain. This required visiting each fast food chain three times. Each time we brought a volunteer and recorded a video of the volunteer eating a meal in the restaurant. Knowing that real videos of obesity patients eating would likely be recorded using low-cost cameras, we simulated the situation by using a Unibrain Fire-I webcam to capture VGA videos. Furthermore, for each food instance, we acquired images and videos under different lighting conditions, with different background, and from different viewing angles. We used a Canon SD1100 digital camera and a Point Grey Bumblebee I stereo camera to obtain rectified and disparity stereo images. We recorded all camera parameters during data acquisition.

Table 1 outlines the data acquisition protocol. While collecting a fast food dataset may seem straightforward, we have learned through practice that the devil is in the details and the details demand meticulous attention.

Per food, Per instance	Restaurant Setting	Laboratory Setting		
Lighting	Ambient	Ambient		
Background	Uncontrolled	White background		
	Tray, table	Turntable covered in white		
Data Collection	Stills	Stills	Stereo images	360 video
Camera Setting	Photo mode, no flash, handheld	Photo mode no flash, 1 second exposure, on tripod	Stereo camera on tripod	Video mode, no flash
Numbers	4 with wrappers, about 90 degrees apart; 4 without wrappers, about 90 degrees apart	6, exactly 60 apart	2 sets of 3 images (1 if round object)	one 5-10 second clip
Extras	1 image of food name tag	1 image of food name tag, 1 reference image of background		
Sizes	2592x1944			VGA

Table 1: Data acquisition protocol.

2.3. Challenges and learning

Mundane details, from getting permission for collecting data in the restaurant, to making sure that volunteers are not visible in images; from not inadvertently pushing the turntable while trying to rotate it, to not forgetting to charge the camera ahead of time, these seemingly *small* things cost much time, money, and inconvenience when overlooked. This is a task that demands painstaking attention to details and a systematic approach. We list more of them below:

- Mount the camera on the tripod at a height where the object is centered, and fills an appropriate portion of the frame;
- For acquiring stereo data, our verification within the context of a normal work desk (1.75 by 0.75m) showed that the distance of 0.75m between the stereo camera and the food gave the smallest estimation error
- Background color and material has significant impact on the quality of images. After some experimentation we found non reflective white background gave the best quality image
- To minimize shadows and highlights, we minimized the use of direct light sources and increased the camera exposure.
- When manually rotating the turntable, our hands should not be visible in the scene

2.4. PFID summary

The PFID collection currently has three instances of 101 fast foods, where each instance of each food has four still images with wrappers and four without wrappers in restaurant environment, six still images in the laboratory setting (without wrappers), two sets of stereo images (left rectified image, right rectified image, disparity image) along the long and short side of the object (one for round objects), and one 360 degree video of the food on a turntable. Figure 1 shows examples of images in PFID

3. BENCHMARKS

We evaluate the accuracy of standard computer vision recognition algorithms on the PFID collection. Specifically, we examine the accuracy with which two popular representations, color histograms and SIFT, are able to capture the image content in our fast food images. The goal is to provide standard baselines for image processing and computer vision researchers who are working in this area rather than to propose such methods as the state of the art.

We employ the following consistent methodology in both of the experiments. Twelve images (from different views of two instances) of each of the 101 food types are utilized as the training set, while the six images (from the third instance) are held out for testing. Each instance is held out in turn and results are averaged over this three-fold cross validation. In particular, we ensure that no instance of a food item ever appears in both the training and test sets. We train a multi-class SVM classifier [16] using the former data using the popular libsvm [10] package, with standard parameters. The following subsections present results for each of the two representations. With 101 classes, the *a priori* recognition rate is below 1%.

3.1. Baseline 1: Color Histogram + SVM Classifier

Color histogram based representations have been popular in object recognition for more than a decade. In this study, we employ a standard RGB 3-dimensional histogram with four quantization levels per channel. Each pixel in the image is mapped to its closest cell in the histogram to generate a 64 dimensional representation for each image.

3.2. Baseline 2: Bag of SIFT Features + SVM Classifier

Bag of features representations have recently become de facto standards as baseline representations in a variety of information retrieval tasks ranging from text classification to content-based image retrieval. The basic idea is to represent each image as a histogram of occurrence frequencies defined over a discrete vocabulary of features and then to use the histogram as a high-dimensional vector in a traditional discriminative framework. In the object recognition community, the SIFT descriptor [9] has emerged as the popular choice for this task, and several studies (e.g., [1]) have demonstrated the merits of building bags of features over a vocabulary of quantized SIFT features.

For a given image, we employ the SIFT interest point detector to identify a sparse set of ``keypoints'' or locations at which descriptors should be computed. These keypoints are localized

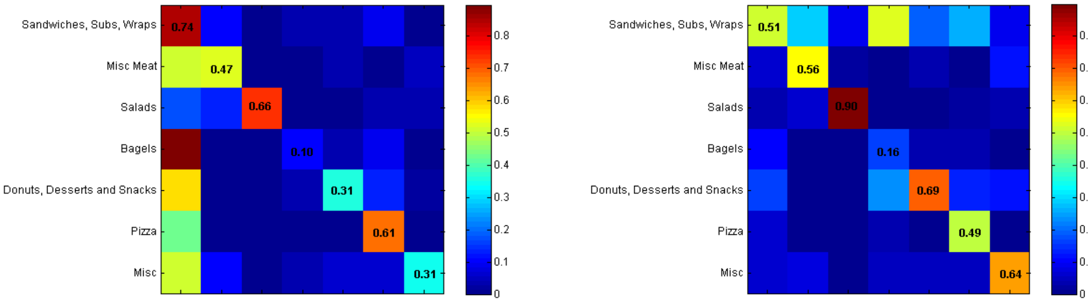

Figure 2: Confusion matrices - color histogram (left) and SIFT (right).

both in scale and in space. For computational efficiency, we run the SIFT detector on images scaled to 20% of their original size and typically find approximately 400 keypoints per image

At each of the keypoints identified by the interest point, we extract a patch and compute its 128-dimensional SIFT descriptor, exactly as described in [9]. These are quantized onto a 1000 word vocabulary using a codebook that was previously generated using K-means clustering. Thus, each feature descriptor is coded by hard assignment to the nearest codebook center, yielding a 1000-dimensional histogram of codeword counts for each image. We normalize the histogram to obtain a ``bag of features'' representation for each image, that is then classified using an SVM as described above.

Figure 3 summarizes the experimental results for the two baselines, averaged over the three-fold cross validation. As can be seen from the overall results, automatic food identification is a challenging task, even when restricted to the apparently simple class of standardized fast foods. The baseline experiments confirm our belief that standard ``out-of-the-box'' approaches to object recognition do not solve the problem and that there is an opportunity for specialized methods to greatly improve accuracy. The approach based on color histogram information alone performs poorly because a food item's aggregated color distribution is not sufficiently discriminative. It does well only on foods with a highly-distinctive color distribution, such as the Pizza Hut dessert pizza, but is otherwise a poor choice. SIFT performs better on the task and generally dominates the color histogram baseline, as expected. It achieves a particularly high accuracy on items such as chocolate chip muffins, due to their distinctive shape and consistent internal appearance. On the other hand, we note that there are many objects where the approach fails.

To better understand the failure modes of the baselines, we generated confusion matrices for the two baseline methods (not shown). We note that there are many food items that are frequently confused with others. In some cases this is to be expected (e.g., sandwiches that differ only in terms of the filling). In other cases, the cause for the misclassifications is less clear. Therefore, we categorized the 101 food items into seven semantically-meaningful categories that correspond to major types of fast food: (1) sandwiches, including subs and wraps; (2) meat preparations, such as fried chicken; (3) salads, typically consisting of greens topped with some meat; (4) bagels; (5) donuts and other sweet snacks; (6) pizza; and (7) miscellaneous category that included a diversity of items such as soup and Mexican-inspired fast food. Figure 2 shows category-level confusion matrices for both baseline algorithms. Note that these 7x7 matrices were not generated by retraining the classification but are merely an alternative visualization of the 101-class results. Due to space limitations, we limit ourselves to a few salient observations. First, we note that for the color histogram method, many food categories (such as bagels) are misclassified as sandwiches. This is probably because the predominant color in a sandwich is that of the bread and this can often match the bread color of the bagel. That effect is not observed in the SIFT baseline, probably because sandwiches and bagels differ in terms of shape. However, it is interesting to see that even a category that seems visually distinctive, such as bagels, is recognized with such low accuracy. Some categories, such as salads, are recognized with high accuracy. However, note that the SIFT algorithm outperforms the color histogram method even in the case of salads, where one would expect color to have significant advantages. From these experiments, we conclude that the bag of SIFT features approach dominates the color histogram method in almost every case and should therefore serve as the preferred baseline for future research.

4. CONCLUSION

We present the first image/video dataset on 101 fast foods. Our intent is to provide a freely available dataset to enable computer vision research on food recognition/classification. We test two standard benchmarks on this dataset, and invite researchers to devise suitable benchmarks and share with the research community. For more details and the complete dataset see [15]. A long term goal is to link PFID to the Food and Nutrient Database for Dietary Studies (FNDDS) database. This will help stimulate other researchers to work on food recognition problems.

ACKNOWLEDGEMENTS

We thank D. Huber, S. Narasimhan, M. Veloso, M. B. Dias, F. Kraus, D. Pomerleau and our volunteers for their advice and contribution. This project was funded by NIH Grant number 1U01HL091736 01 and Intel Labs Pittsburgh.

REFERENCES

[1] Leibe et al., Combined object categorization and segmentation with an implicit shape model, Workshop on Statistical Learning Computer Vision, 2004.

[2] Wu et al., Fast Food Recognition from Videos of Eating for Calorie Estimation, ICME, 2009.

[3] Li et al., Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Workshop on Generative-Model Based Vision. 2004.

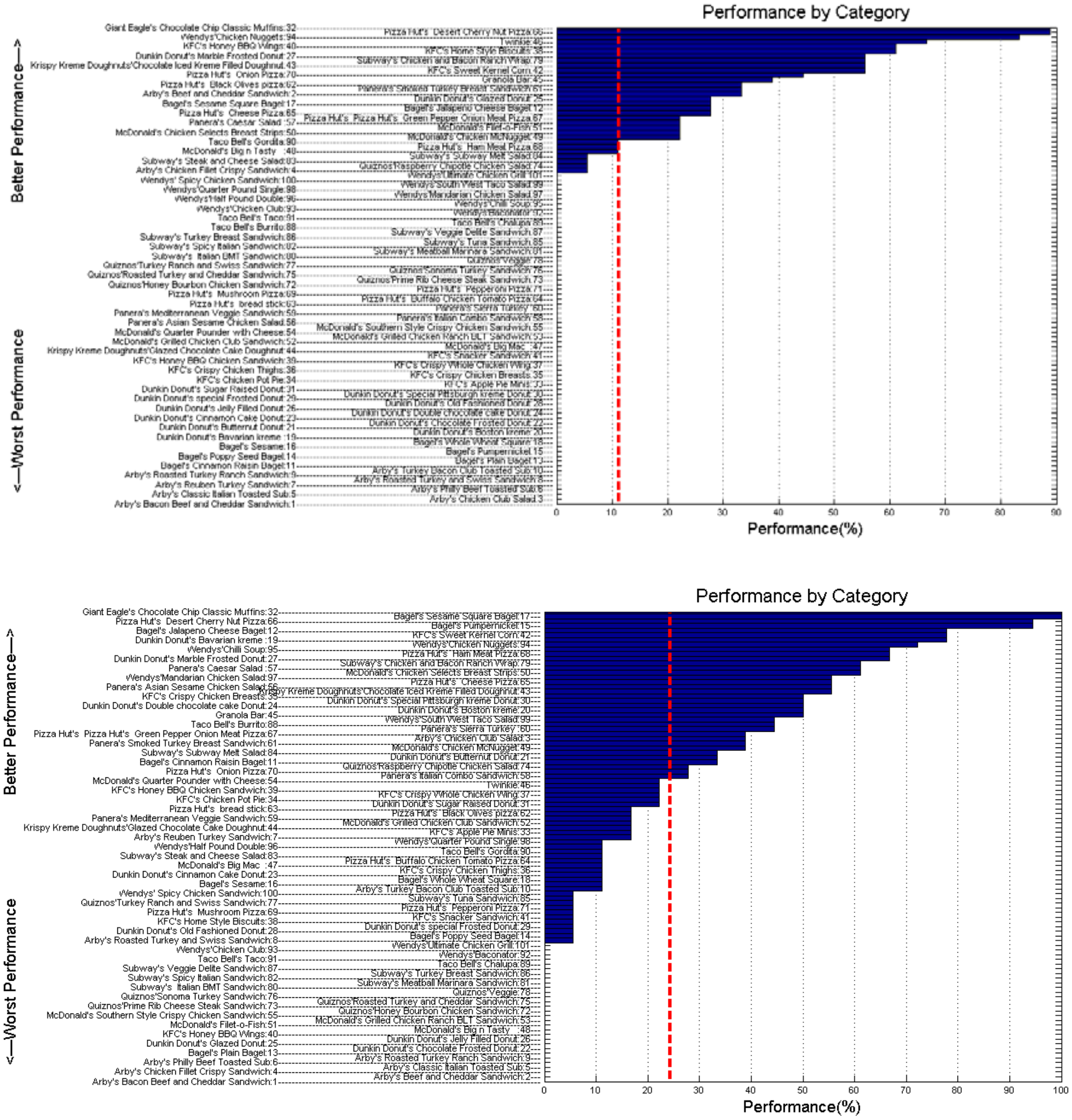

Figure 3: Performance by category - color histogram (top) and SIFT (bottom).

[4] M. Everingham,et al. The 2005 PASCAL Visual Object Classes Challenge. In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment., LNAI 3944, 2006.

[5] M. Everingham, et al., A. Zisserman, The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascalnetwork.org/challenges/VOC/voc2008/workshop

[6] P. Phillips, et al., The FERET evaluation methodology for face recognition algorithms. IEEE Trans. PAMI 22 (10), 2000.

[7] N. Yao et al., A Video Processing Approach to the Study of Obesity, Proceedings of ICME, 2007.

[8] Yang et al., Layered Object Categorization, ICPR, 2008.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 60 (2), 2004.,

[10] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw /~cjlin/libsvm/

[11] National Institutes of Health Genes GEI Exposure Biology Program. http://www.gei.nih.gov/exposurebiology/.

[12] http://www.cdc.gov/nccdphp/dnpa/obesity/trend/maps/

[13] G. Griffin, et al., Caltech-256, Caltech Tech Report, 2007.

[14] Food Surveys Research Group: Agricultural Research Service. USDA Food and Nutrient Database for Dietary Studies, 3.0,

[15] M. Chen, et al.. PFID: Pittsburgh Fast-Food Image Dataset. Carnegie Mellon Technical Report. CMU TR-06-09, 2009.

[16] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[17] Chen et al., Food safety inspection using “from presence to classification” object-detection model, Pattern Recognition, 2001.